(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号

特開平5－265829

(43)公開日　平成５年(1993)10月15日

| (51)Int.Cl.[5] | | 識別記号 | 庁内整理番号 | ＦＩ | 技術表示箇所 |
|---|---|---|---|---|---|
| Ｇ０６Ｆ | 12/00 | ５３３　Ｊ | 7232－5B | | |
| | 11/20 | ３１０　Ｅ | 7832－5B | | |
| | 12/00 | ５３１　Ｄ | 7232－5B | | |
| | | ５４５　Ａ | 7232－5B | | |

審査請求　未請求　請求項の数８(全 11 頁)

(21)出願番号　特願平4－90248

(22)出願日　平成４年(1992)３月16日

(71)出願人　000005108
株式会社日立製作所
東京都千代田区神田駿河台四丁目６番地

(72)発明者　櫻庭　健年
神奈川県川崎市麻生区王禅寺1099番地　株式会社日立製作所システム開発研究所内

(72)発明者　福澤　淳二
神奈川県川崎市麻生区王禅寺1099番地　株式会社日立製作所システム開発研究所内

(72)発明者　黒田　澤希
神奈川県川崎市麻生区王禅寺1099番地　株式会社日立製作所システム開発研究所内

(74)代理人　弁理士　笹岡　茂　(外１名)

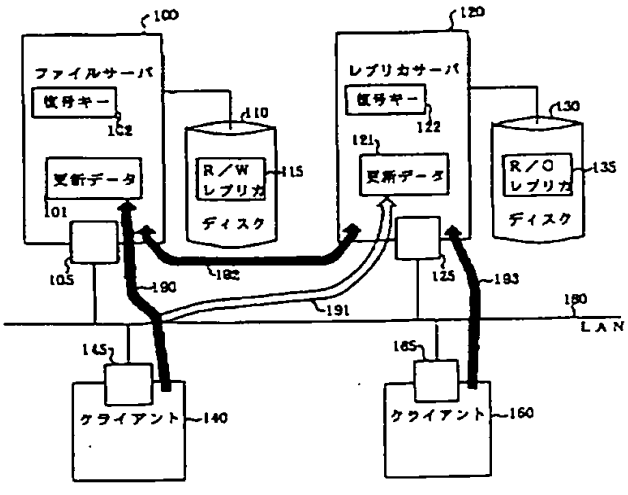
(54)【発明の名称】　レプリケートファイル更新方式

(57)【要約】

【目的】　分散データ処理システムにおける、レプリカサーバの読み出し専用レプリカを、新たに更新データの送信を行なうことなく効率よく最新化する。

【構成】　ファイルサーバ１００にある、或るファイルの書き込み可能レプリカ１１５の更新を、通信ネットワーク１８０に接続されたクライアント１４０が通信路１９０を通じて行う時、該クライアントと前記ファイルの読み出し専用１３５を持つレプリカサーバ１２０との間の擬似的な通信路１９１を用いて同時にレプリカサーバで更新データを取り込み、該更新データを読み出し専用レプリカの最新化処理に用いることにより、当該最新化処理ではファイルサーバからレプリカサーバへの更新データ送信を不要とする。更にファイルサーバとレプリカサーバとの間ではレプリカサーバが取り込んだ更新情報の妥当性を確認するための通信を行って、レプリカ１１５、と１３５の同一性を保証する。

【図１】



1

【特許請求の範囲】

【請求項１】　通信ネットワークと、

当該通信ネットワークに接続された少なくとも１つのファイルサーバと少なくとも１つのレプリカサーバと、

前記通信ネットワークに接続され前記ファイルサーバあるいはレプリカサーバの有するレプリケートファイルにアクセス可能な少なくとも１つのデータ処理システム（クライアント）とからなる分散データ処理システムにおけるレプリケートファイル更新方式であって、

前記レプリカサーバは、前記通信ネットワークを介してレプリカサーバに対して行なわれる前記クライアントからのアクセスおよび該クライアントと前記ファイルサーバの間で行われる該ファイルサーバのレプリケートファイル更新のための前記通信ネットワーク上の交信を監視し、前記アクセスの検知および更新されるファイルサーバ内のレプリケートファイルと同一のファイルをレプリカサーバが有するか否かの検知をする検知手段と、

該検知手段により同一のファイルを有するという検知結果が得られた場合前記クライアントからファイルサーバに送信される更新内容を取り込む手段を備え、

前記レプリカサーバは、取り込んだ更新内容によりレプリケートファイルの更新を行うようにしたことを特徴とするレプリケートファイル更新方式。

【請求項２】　請求項１記載のレプリケートファイル更新方式において、

前記ファイルサーバとレプリカサーバは、取り込んだ更新内容によるファイル更新が何回目のファイル更新であるかを示すファイル更新回数を得る手段を備え、

前記レプリカサーバは、得られたファイル更新回数と前記ファイルサーバから送信されたファイル更新回数とを比較する更新回数比較手段を備え、

前記レプリカサーバは、前記更新回数比較手段による比較の結果、ファイル更新回数が一致したとき取り込んだ更新内容によりレプリケートファイルの更新を行うようにしたことを特徴とするレプリケートファイル更新方式。

【請求項３】　請求項２記載のレプリケートファイル更新方式において、

前記レプリカサーバは、前記更新回数比較手段による比較の結果、ファイル更新回数が一致しないとき、更新内容送信要求を前記ファイルサーバに送信し、該ファイルサーバから送信された更新内容を受信し、受信した更新内容によりレプリケートファイルの更新を行うようにしたことを特徴とするレプリケートファイル更新方式。

【請求項４】　請求項２記載のレプリケートファイル更新方式において、

前記ファイルサーバとレプリカサーバは、取り込んだ更新内容の妥当性を示す妥当性判定用データを生成する生成手段と、

前記レプリカサーバは、得られた妥当性判定用データと

2

前記ファイルサーバから送信された妥当性判定用データとを比較する妥当性比較手段を備え、

前記レプリカサーバは、前記妥当性比較手段による比較の結果、妥当性判定用データが一致したとき取り込んだ更新内容によりレプリケートファイルの更新を行うようにしたことを特徴とするレプリケートファイル更新方式。

【請求項５】　請求項４記載のレプリケートファイル更新方式において、

前記レプリカサーバは、前記更新回数比較手段による比較の結果と妥当性比較手段による比較の結果の両方またはいずれか一方が不一致のとき、更新内容送信要求を前記ファイルサーバに送信し、該ファイルサーバから送信された更新内容を受信し、受信した更新内容によりレプリケートファイルの更新を行うようにしたことを特徴とするレプリケートファイル更新方式。

【請求項６】　請求項１乃至請求項５のいずれかの請求項記載のレプリケートファイル更新方式において、

前記検知手段を、前記アクセスの検知をする第１の検知手段と前記更新されるファイルサーバ内のレプリケートファイルと同一のファイルをレプリカサーバが有するか否かの検知をする第２の検知手段に分割して独立に設けたことを特徴とするレプリケートファイル更新方式。

【請求項７】　請求項６記載のレプリケートファイル更新方式において、

前記レプリカサーバに少なくとも２つのプロセッサを設け、前記クライアントからのアクセスに対する処理と、前記更新されるファイルサーバ内のレプリケートファイルと同一のファイルに対するレプリカサーバにおける処理を夫々別々のプロセッサにより行なうようにしたことを特徴とするレプリケートファイル更新方式。

【請求項８】　請求項１乃至請求項７のいずれかの請求項記載のレプリケートファイル更新方式において、

前記レプリカサーバは、前記クライアントから前記ファイルサーバに送信される暗合化された情報を解読するのに必要な暗合複合情報を格納する格納手段を設け、予め前記ファイルサーバから前記レプリカサーバに送信される暗合複合情報を該格納手段に格納し、前記クライアントから前記ファイルサーバに送信される更新内容を取り込むとき前記暗合複合情報を用いて取り込むようにしたことを特徴とするレプリケートファイル更新方式。

【発明の詳細な説明】

【０００１】

【産業上の利用分野】本発明は分散データ処理システムにおけるレプリケートファイルの効率の良い運用に好適なレプリケートファイル更新方式に関する。

【０００２】

【従来の技術】

（１）スヌープキャッシュ

スヌープキャッシュについては、例えば、丸善株式会

社、ＭＡＴ、電子情報通信編、並列処理機構（１９８９年）、ＰＰ１６７－２０７に記載されている。

【０００３】各プロセッサが別々のキャッシュメモリを備え、かつ共通の主記憶にアクセスするように構成されたマルチプロセッサシステムでは、主記憶上の同一のアドレスにあるデータを各プロセッサのキャッシュメモリにコピーしてアクセスする。この場合、あるプロセッサが主記憶を更新すると、更新したプロセッサ以外のプロセッサが、そのキャッシュメモリに更新されたアドレスのデータを持っているときは、そのデータを無効化するなどして、キャッシュ間の無矛盾性（コヒーレンシー）を保証する必要がある。スヌープキャッシュはそのためのキャッシュ制御方式、ないしその制御方式を採用するキャッシュメモリのことである。

【０００４】スヌープキャッシュでは各プロセッサがバス結合されているマルチプロセッサシステムにおいて、１つのプロセッサが主記憶を更新すると、各プロセッサのキャッシュ制御回路はバスを監視することによって更新アドレスを知ることができ、該当するデータをキャッシュメモリに持っていればそれを無効化してコヒーレンシーを保つ。スヌープキャッシュでは各プロセッサが監視するのは、それと同等の機能を持つ他のプロセッサのメモリへの書き込みである。上記文献ＰＰ１９６－１９７では、書き込み放送型のキャッシュ更新プロトコルについて解説している。このプロトコルでは他のプロセッサがメモリ更新するデータが自らのキャッシュ内にある場合、そのキャッシュを無効化するのではなく、データを該当するキャッシュに直ちに書き込む。スヌープキャッシュ適用システムではバスの信頼性が高く、１プロセッサによるメモリ更新は他のプロセッサに確実に伝えられ、各スヌープキャッシュにおけるコヒーレンシーの保証動作が確実に起動される。

【０００５】（２）分散ファイルシステム

分散データ処理システムは、ＬＡＮ（Ｌｏｃａｌ　Ａｒｅａ　Ｎｅｔｗｏｒｋ）と呼ばれる通信ネットワークにパーソナルコンピュータやワークステーションなどのデータ処理システムを複数接続し、相互に通信ネットワークを介して情報の授受を行いながら、それぞれの処理を遂行するシステムである。分散データ処理システムにおける重要な技術として、通信ネットワーク上の１つのデータ処理システムから他のデータ処理システムに存在するファイルのアクセスを可能とする分散ファイルシステムがある。分散ファイルシステムは、１つのデータ処理システムに存在するファイルを複数のデータ処理システムが共用することを可能とする。ここでは、共用されるファイルが存在するデータ処理システムをファイルサーバと呼び、ファイルサーバ上の共用ファイルにアクセスするデータ処理システムをクライアントと呼ぶことにする。

【０００６】特定の共用ファイルが頻繁にアクセスされ



るとそのファイルサーバに要求が集中し、分散データ処理システム全体の性能上のボトルネックとなる可能性がある。そこでこのファイルのコピーを第２のファイルサーバにも置き、このファイルへのアクセス要求の一部を第２のファイルサーバで処理することにより、２つのファイルサーバの間で負荷分散し、前記のボトルネックの解消を図ることが行われる。同一ファイルのコピーを複数のデータ処理システムに置くことにより、ファイル破壊が発生してもコピーを代替として用いることにより、当該ファイルのへのアクセスが中断せず、また回復も容易となるので、分散データ処理システム全体の信頼性向上にもなる。同一ファイルのコピーの各々をそのファイルのレプリカ、あるいはレプリケートファイルと呼び、コピーを作成することレプリケーションという。

【０００７】分散ファイルシステムについてはエー・シー・エム、コンピューティングサーベイズ、第２２巻、第４号（１９９０年）ＰＰ３２１－３７４（ＡＣＭ，Ｃｏｍｐｕｔｉｎｇ　Ｓｕｒｖｅｙｓ，Ｖｏｌ．２２，Ｎｏ．４（１９９０）ＰＰ３２１－３７４）において解説されており、レプリケーションについては同文献のＰＰ３３９－３４０に論じられている。上記文献にもある通り、レプリケーションでは複数あるレプリカの更新方式が問題である。更新が発生した時点で全てのレプリカに同様の更新を行う方式はレプリカ間の同等性の保証はできるがオーバヘッドが大きい。一方、レプリカ間の同等性を全く保証しない方式は可用性、及び性能の観点からは良好であるが、同等性がくずれるため、処理結果の正当性が保証されない。よく行われる方法は書き込み可能なレプリカ１つと読み出し専用のレプリカを複数持ち、書き込み可能レプリカに対する更新内容を読み出し専用レプリカのあるデータ処理システムに後から転送して、読み出し専用レプリカを最新化する、即ち書き込み可能レプリカの最新状態を読み出し専用レプリカにアトミック性を保証しながら反映させる方法である。一時的にレプリカ間の同等性がくずれるが、適当な時間経過の後は同等性が回復され、ファイル使用中はファイル内部の一貫性が保証されるので使用に耐え得る方式となっている。特に読み出し専用ファイル、ないし更新が極めてまれなファイルに対して好適である。

【０００８】以下では記述の簡単化のために書き込み可能レプリカが存在するファイルサーバを単に（そのファイルの）ファイルサーバと呼び、読み出し専用レプリカが存在するファイルサーバを（そのファイルの）レプリカサーバと呼ぶことにする。ファイルサーバ、及びレプリカサーバという呼び方は特定のファイルを決めたときに定まる概念であり、物理的に同じデータ処理システムが特定ファイルのファイルサーバであると同時に、別のファイルのレプリカサーバであることは可能である。

【０００９】

【発明が解決しようとする課題】書き込みが頻繁なファ

イルの場合は、読み出し専用レプリカの最新化処理が頻発し、最新化処理に伴う更新データ転送量が多くなる。そのため、書き込み可能レプリカを持つファイルサーバでは更新ファイル、ないし更新データの送信処理のオーバヘッドが発生し、また通信路に新たな負荷が発生してそのビジー率が上がる。１つのファイルに対し、複数のレプリカを持った場合には、持たない場合に較べ、これらのオーバヘッドはレプリカの個数倍だけ増大する。

【００１０】ファイルサーバのオーバヘッドは複数のファイルサーバを導入して負荷分散することにより緩和できる。しかし、通信路は分散データ処理システムに唯一の資源であり、その負荷の緩和は容易ではない。分散データ処理システムを評価する１つの尺度に規模拡張性（スケーラビリティ）があるが、通信路はそのボトルネックの１つである。即ち、レプリケーションによるファイルの可用性、及び信頼性の向上とシステムの規模拡張性の間にはトレードオフが存在し、その解決方式が問題である。本発明の目的は、ファイルサーバにおけるレプリカ更新のための更新データ送信オーバヘッド、及び通信路の負荷を発生させることのないレプリケートファイルの効率の良い、かつ高信頼な更新方式を提供することにある。

【００１１】

【課題を解決するための手段】上記目的を達成するために、本発明では書き込み可能レプリカの更新に伴う、ファイルサーバとクライアントとの通信を、読み出し専用レプリカを持つデータ処理システムであるレプリカサーバから監視し、レプリカサーバは書き込みを認知したら書き込み内容を取り込んで読み出し専用レプリカの最新化処理を行う。また、レプリカサーバの読み出し専用レプリカの更新に際して、更新内容の妥当性を更新対象ファイルの更新回数あるいは妥当性判定用データにより判定し、この判定を、レプリカサーバで得られた更新対象ファイルの更新回数あるいは妥当性判定用データとファイルサーバで得られた更新対象ファイルの更新回数あるいは妥当性判定用データとの一致をファイルサーバとの通信により確認することにより行なう。レプリカサーバは、妥当性が確認された場合は、読み出し専用レプリカの最新化、即ち更新を行い、エラーが検出された場合は、ファイルサーバから改めて最新状態のレプリカを送信させ、これにより読み出し専用レプリカを更新する。

【００１２】

【作用】レプリカサーバは、読み出し専用レプリカの最新化のための更新データとして、ファイルサーバにおける書き込み可能レプリカの更新処理時に授受された更新情報を取り込み、これを用いて読み出し専用レプリカを更新する。このため、読み出し専用レプリカの最新化のために新たにファイルサーバからレプリカサーバに更新情報を送る必要がなく、従って通信ネットワークの負荷を抑えることができる。また、読み出し専用レプリカの



最新化に必要な更新情報の特徴をよく表し、かつ更新情報全体よりはるかに短いデータである更新対象ファイルの更新回数あるいは妥当性判定用データをファイルサーバから受け取って、あるいはファイルサーバに送って、両サーバで得られた更新回数あるいは妥当性判定用データの一致を調べ、データの妥当性を確認する。これにより、通信ネットワーク等で生じるデータ誤りに対処できる。妥当性が確認された場合は上記同様、新たな更新情報の送信を行わずに読み出し専用レプリカの最新化を実行する。確認に用いるデータは極めて小さいので通信ネットワークの負荷の増加は問題にならない。妥当性確認の結果、更新データの不当性（エラー）が認められた場合は、書き込み可能レプリカの全内容、ないし最新化に必要な更新情報をファイルサーバからレプリカサーバに送信して書き込み専用レプリカの最新化を実行する。この場合はファイルサーバ、及び通信ネットワークにおける負荷の緩和は実現できないが、エラーの発生率は極めて小さいと考えられるため、十分な負荷削減効果があり、かつ本発明を適用しない場合と同程度の信頼性を確保することができる。

【００１３】

【実施例】以下、図を用いて本発明の実施例を説明する。図１は本発明が適用可能な分散データ処理システムの概要を示した構成図である。書き込み可能なレプリケートファイル（レプリカ）を備えるデータ処理システムであるファイルサーバ１００、読み出し専用のレプリケートファイルを備えるデータ処理システムであるレプリカサーバ１２０、第１のクライアント（ファイルサーバあるいはレプリカサーバの有するレプリケートファイルにアクセス可能なデータ処理システム）１４０、第２のクライアント１６０はそれぞれ独立したデータ処理システムであり、各データ処理システムは共通の通信ネットワーク１８０にそれぞれインターフェースボード１０５、１２５、１４５、及び１６５により接続されており、相互に通信、及びデータ転送が可能である。例えばファイルサーバ１００、及びレプリカサーバ１２０は高性能なワークステーション（ＷＳ）、第１、第２のクライアントは性能は並みで安価なＷＳ、あるいはパーソナルコンピュータ（ＰＣ）からなり、通信ネットワークとしてローカルエリアネットワーク（ＬＡＮ）を使用するのが典型的な構成の例である。ファイルサーバ１００、及びレプリカサーバ１２０はそれぞれファイル格納のためにディスク装置１１０、及び１３０を有しており、そこに格納されたファイルは、クライアント１４０、１６０から共通にアクセスすることができる。特定のファイルを固定して考えるとき、そのファイルの書き込み可能（Ｒ／Ｗ）レプリカ１１５はファイルサーバ１００のディスク１１０に格納され、そのファイルの読み出し専用（リードオンリ：Ｒ／Ｏ）レプリカ１３５はレプリカサーバ１２０のディスク１３０に格納されている。クライ

アント１４０はこのファイルをファイルサーバ１００上のレプリカ１１５を読み出し、クライアント１６０は同じファイルをレプリカサーバ１２０上のレプリカ１３５を読み出すことにより、ファイルサーバとレプリカサーバにファイル転送処理オーバヘッドを分散させることができる。

【００１４】レプリカ１１５は書き込み可能レプリカであり、クライアント１４０はこれの更新を要求することができる。更新情報１０１、即ちファイルの更新後の新しいデータはクライアント１４０から物理的な通信路１４５、１８０、及び１０５から構成される通信路１９０を経て伝達される。ファイルサーバ１００では更新要求に従い、このファイルの書き込み可能レプリカ１１５を更新する。この時点でレプリカ１１５とレプリカ１３５の間には内容的な不一致が生じており、将来、適当な時点に何らかの方法で両者を一致させる必要がある。

【００１５】通信ネットワークを介した通信は送信側、及び受信側の交信手順を定めたプロトコルに乗っ取って行われ、この中には受信すべきデータ処理装置の指定方法、即ちアドレスの指定方法も含んでいる。データ転送自体は通信路上に電気的な振動を発生させ、あるいはそれを検出することにより行われ、通信は、各データ処理システムが検出した電気振動に含まれるアドレス情報を読み取って、自らへの通信であるか否かを判定することにより成立する。

【００１６】本実施例ではレプリカサーバ１２０のインターフェースボード１２５においてレプリカサーバ１２０への本来の通信だけでなく、ファイルサーバ１００への通信も取り込み、書き込み可能レプリカの更新とその更新内容を知る。これは通信ネットワーク１８０上の電気振動はインターフェースボード１０５と同様、インターフェースボード１２５においても検出されるので可能である。即ち、クライアント１４０からレプリカサーバに到る通信路１９１が事実上あって、これを通じてファイルサーバへの更新データの送信を監視する事を可能とし、得られた更新情報１２１を読み出し専用レプリカ１３５の最新化に使用する。

【００１７】分散データ処理システムでは通信のセキュリティのために通信内容を暗合化することがある。ファイルサーバ１００はその複合キー１０２をあらかじめ知っているので交信データ１０１の交信を行う事ができるが、レプリカサーバではデータを取り込んでも、そのままではそれを正しく更新情報として使用することはできない。そこでファイルサーバ１００はレプリカサーバ１２０にあらかじめ複合キー１０２を送信しておく必要がある。レプリカサーバ１２０では当該複合キーを自らのシステム内に複合キー１２２として保持する。取り込んだ更新データ１２１の複合にはこの複合キー１２２を用いて正しくレプリカの更新を行うことができる。

【００１８】図２はレプリカサーバ１２０におけるファイルサーバの通信を取り込むためのインターフェースボード２３０をレプリカサーバ１２０への通信を受け付ける本来のインターフェースボード２１０の他に設けて、より確実な通信監視を行うようにした他の実施例を示している。

【００１９】通信ネットワーク１８０とのインタフェースであるインターフェースボード２１０は電気的接続の他に、通信のあて先アドレスの処理を行う。即ち、指定アドレスが自らのアドレスであった場合は、通信データを取り込み、その処理をデータ処理システム１２０に行わせるためにデータ処理システム１２０のプロセッサ２６０に割込み２８０をかける。通信データの処理中にファイルサーバで更新が行われると、レプリカサーバ１２０のプロセッサ２６０はその更新データの取り込みに失敗する可能性がある。そこで、ファイルサーバ１００の通信を監視するために、レプリカサーバ１２０本来のインターフェースボード２１０とは別に、ファイルサーバ１００の通信監視専用のインターフェースボード２３０を設ける。さらにインターフェースボード２３０が取り込んだデータを確実に処理するため、専用のプロセッサ２５０を設け、監視データの処理の場合は常にプロセッサ２５０に割り込むように構成すると、より良好な性能を実現できる。クライアント１４０がファイルサーバ１００と交信すると事実上、物理的な通信路１４５、１８０、及び２３０からなる通信路２９１が構成される。これは図１の１９１の通信路に相当するものである。

【００２０】プロセッサ２５０とプロセッサ２６０の関係は以下のようである。すなわち、プロセッサ２５０がファイルサーバにおけるファイルの書き込み可能レプリカ更新を監視し、その更新内容を取り込んでいる間はプロセッサ２６０はレプリカを用いたサービスを含む一般の処理を実行し続けることができる。読み出し専用レプリカのバージョンアップ処理では、分散ファイルシステムの制御方針にもよるが、新旧レプリカの切り替え処理とファイルサーバにおけるファイルの更新が並行して実行される場合がある。この更新はレプリカサーバにおいて作成中の新レプリカの次のバージョンに反映されるべきものであり、更新内容の取り込みを引続き行なう必要がある。更新内容取り込み専用プロセッサを用いた複数のプロセッサによって構成されるレプリカサーバの場合、レプリカのバージョンアップ処理は一般処理用プロセッサ２６０が新レプリカを作成し、プロセッサ２５０はレプリカ更新監視を続行する。そのため、バージョンアップ処理が起動される段階で、更新データバッファの切り替えと同時にそれまでの更新データバッファのプロセッサ２５０からプロセッサ２６０への引継ぎが行なわれ、それに伴う、両プロセッサ間での同期処理が必要となる。

【００２１】次にレプリカサーバ１２０でのファイル交信データの処理について説明する。レプリカサーバ１２

０のインターフェースボード２１０、２３０は通信ネットワーク１８０で通信が行われると通信データを取り込むと同時にレプリカサーバ１２０のプロセッサに割込みをかけてその処理を要求する。図３はこのときの割込み処理を示している。割込み処理３００では、ステップ３１０にて前記取り込まれたデータの内容、ないし割込み情報から、当該データが、当レプリカサーバに読み出し専用レプリカがあるファイルに係わるものでありかつ書き込み可能レプリカを持ついずれかのファイルサーバとの通信に係わるものであるか否かを調べる。該当するファイルサーバの通信である場合は、ステップ３２０に進む。

【００２２】ステップ３２０ではその通信内容がファイル更新に係わるものか否かを調べ、ファイル更新用の通信ならば、ステップ３３０に進む。ステップ３３０では更新対象であるファイルが、当レプリカサーバに読み出し専用レプリカがあるファイルであるか否かを調べる。もし、レプリカが存在するファイルならばステップ３５０に進む。ステップ３５０では前記データをレプリカ更新データとして取り込み、保存する。更にステップ３６０にて、レプリカ、及びレプリカ更新データの管理情報を作成、ないし更新し、後の読み出し専用レプリカの最新化処理で用いる。

【００２３】次にステップ３１０、３２０、及び３３０の判定処理で用いるデータについて説明する。図４は割込み時にメモリ上に取り込まれている通信データの構造の例を示している。当該データはそれを参照する処理のレベルに対応して入れ子構造になっている。データ構造４００は通信ネットワーク上を伝送される時のデータ構造である。発信元アドレス４０１、あて先アドレス４０２、データタイプ４０３、エラー検出用冗長データ４０５、及びデータ本体を含んでいる。

【００２４】ステップ３１０では、あて先アドレス４０２を参照することにより、当該通信データが関係するファイルサーバへの通信であるか否かを判定できる。データ構造４００のデータ本体４０４は上位レベルの処理で用いるデータ構造４２０を含んでいる。データ構造４２０は発信元ＩＤ４２１、あて先ＩＤ４２２、制御情報４２３、及び上位処理に渡すデータ４２４を含む。あて先ＩＤ４２２はその通信を受け取るべき機能、ないしプログラムを表す。レプリカサーバ１２０はファイルサーバ１００の制御プログラムがどのようなあて先ＩＤを持つかを知っており、このフィールド４２２により、ファイルサーバに対する通信であることが分かる。

【００２５】データ４２４は４４０に示すデータ構造を持つ。サービスＩＤ４４１にはファイルサーバの制御プログラムに対するサービス要求内容が表示される。当該フィールドを参照することにより、ファイルの更新処理か否かが分かる。従って、ステップ３２０ではフィールド４２２、４４１を参照してファイル更新要求であることを判定する。ファイル更新要求である場合はデータ構造４４０の後続のフィールドには、更新対象となるファイルのファイルＩＤ４４２、更新位置の相対アドレス４４３、更新範囲を表す更新長４４４、及び更新内容そのものである更新データなどがある。レプリカサーバでは格納しているレプリカの管理テーブルを所有しており、ステップ３３０ではファイルＩＤ４４２が示すファイルのレプリカを持っているか否かを判定することができる。更新内容を表すフィールド４４３、４４４、４４５はステップ３５０で用いられ、内容の一部はディスクに格納される。



【００２６】次に更新データの妥当性について説明する。レプリカの更新データはレプリカサーバ１２０とファイルサーバ１００で別々に取り込まれ、更新に使われる。特にファイルサーバ１００では更新要求をしたクライアント１４０との間で適当なプロトコルに乗っ取り、確実なデータ転送が行われるが、レプリカサーバではこの交信を監視するのみなので、取り入れた更新データの妥当性は必ずしも保証できない。データを取り入れることができた場合には、データにはエラー検出用の冗長データが付加されており、データの妥当性はある程度保証できる。しかし、例えばレプリカサーバが過負荷のため、ファイルサーバ１００の通信の監視にもれが生じ、更新データが失なわれる可能性がある。従って、何らかの手段により、レプリカサーバが取り込んだ更新データの妥当性チェックを、ファイルサーバ１００を交えて行う方が望ましい。

【００２７】各取り込みデータのエラー検出用冗長データの有効性を仮定すれば、毎回の当該データのチェックの他に、ファイルサーバ、及びレプリカサーバにおける更新回数の一致を確認すればレプリカサーバにおいて取り込んだ更新データの妥当性は保証されたと考えることができる。すなわち、例えばファイルサーバ及びレプリカサーバに格納されているファイルＡの或る時点以降の更新について見ると、この時点以降の最初の更新を１回目の更新とし、以下更新が或る度に２回目、３回目、……、ｎ回目の更新とする。各回の更新はファイルＡ全体の更新であってもよく、また、１回目の更新ではファイルＡの前半部分の２０バイト、２回目の更新ではファイルＡの後半部分の３０バイト、、３回目の更新ではファイルＡの中央部分の４０バイトというようなファイルＡの部分的更新であってもよい。

【００２８】一方、エラー検出用冗長データの有効性が不十分であったり、回数の一致だけでは妥当性に不安がある場合には、更新データに依存した妥当性判定用のデータをレプリカサーバ、及びファイルサーバの双方で作成し、両者の一致を確認する必要がある。妥当性判定のための交信は更新されたデータよりはるかに小さいことが、本発明の目的である通信ネットワークの負荷を抑えるための条件である。妥当性判定用データとして、各回

の更新データのエラー検出用冗長データを送り、両者を突き合わせる。これにより、不一致を検出するだけでなく、不一致の場合には、エラー検出用冗長データの不一致が発生した回の更新データのみをレプリカサーバに送るようにすれば、全ての更新データを送る場合に比べて、データ転送量を削減することができる。

【００２９】図５は、この妥当性判定用データの構造を示している。妥当性判定用データ５００は、前回のバージョンアップ以降から数えて１回目のから最新の更新までの更新データに付随するエラー検出用冗長データフィールド５１０、５２０、……、５３０からなる。例えば２回目の更新時の更新データのみをレプリカサーバに転送するのみでレプリカの同一性を保証することができる。また、別の妥当正判定の仕方として、ファイルサーバ及びレプリカサーバの双方でｎ回目の更新が行なわれた後の更新内容についてのエラー検出用冗長データを生成し、エラー検出用冗長データのみをファイルサーバからレプリカサーバに判定用データとして転送し、その一致を確かめるようにしてもよい。これによれば、途中でエラーが発生しても、また更新データの取り込みが欠落しても、ファイルサーバからｎ回目までの更新データが送られ書き換えが行なわれることにより、データの上書きが行なわれ、最終的に内容が一致すればレプリカとして有効であるため、エラーのペナルティを軽減する効果がある。特に、この場合、更新回数を更新データの妥当性判定のためのデータとして用いなければならないという条件を外すことができる。

【００３０】次にファイルサーバで用いられるレプリカ管理テーブルについて説明する。図６はファイル管理エントリ６００とそのレプリカを管理するレプリカ管理エントリ６２０の構造を示している。管理対象となるファイル、及びレプリカは複数存在するのでいずれもチェイン構造を伴っている。ファイル管理エントリ６００（ＦＭＥ）には当該エントリが管理するファイルのファイルＩＤ６０１、当該ファイルのレプリカのレプリカ管理エントリ（ＲＭＥ）へのポインタ６０３を含む。レプリカ管理エントリ６２０にはレプリカＩＤ６２１、当該レプリカが存在するデータ処理装置のアドレス６２３、当該レプリカの現在の状態を表すカレントバージョン６２４、当該バージョンをレプリケートしてから後の当該ファイルの更新回数を示す更新カウント６２５、更に更新経過を特徴づける妥当性判定用データ６２６を含む。

【００３１】次にレプリカサーバにおける読み出し専用レプリカの最新化処理について説明する。図７はレプリカサーバにおける読み出し専用レプリカの最新化処理の１実施例におけるファイルサーバにおける処理７００、及びレプリカサーバにおける処理７５０を示すフローチャートである。本図には両サーバ間の通信を表すかぎ付き矢印７１１、７３１、７７１、及び７７２が示されている。また通信７１１で送信されるデータ７２０も示されている。本実施例では、読み出し専用レプリカの最新化処理はファイルサーバから起動するものとする。



【００３２】まず、ファイルサーバがステップ７１０にてレプリカサーバにその最新化要求を発行する（７１１）。この通信７１１では送信データ７２０として、最新化すべきレプリカのファイルＩＤ７２１、前回最新化してから後の、当該ファイルの書き込み可能レプリカの更新回数７２２、更に必要ならば更新データの妥当性判定用データ７２３などが送られる。更新回数７２２、及び判定用データ７２３はファイルサーバにおけるレプリカ管理エントリ６２０内の更新カウント６２５、及び判定用データ６２６の値が用いられる。本通信７１１により、レプリカサーバにおけるレプリカ最新化処理７５０が起動される。レプリカサーバではステップ７５５において送信された更新回数７２２と判定用データ７２３とレプリカサーバ内部で計数、あるいは生成した更新回数、及び判定用データとを比較して、レプリカサーバにおける更新データの妥当性を評価する。妥当性が確認されたならば、ステップ７６０に進む。

【００３３】ステップ７６０ではレプリカサーバが取り込んだ更新データを用いて最新化処理を当該レプリカサーバ内の読み出し専用レプリカに対して行う。最新化処理が成功した場合はステップ７６２でファイルサーバからの最新データの送信が不要であることを送信７７１し、ステップ７８０にてレプリカ管理情報を更新してレプリカ最新化処理を終了する。また最新化処理が失敗した場合はステップ７７０に進む。ステップ７５５におけるレプリカサーバが取り込んだ更新データにの妥当性が確認されなかった場合、及びステップ７６１で最新化処理が失敗したと判断された場合は、ステップ７７０にて、ファイルサーバに対して指定ファイルの最新版の送信を要求する応答７７２を返す。

【００３４】ファイルサーバはステップ７１５でレプリカサーバからの応答７７１、７７２を調べ、最新バージョンの送信要求でない、即ちレプリカサーバにおけるレプリカの最新化が完了した場合はステップ７４０でレプリカ管理情報の更新を行い、処理を終了する。また、ステップ７１５で最新バージョンの送信が要求されていることが分かった場合には、ステップ７３０に進み、最新バージョンの送信７３１を行う。処理は再びステップ７２０に戻り、レプリカサーバにおけるレプリカの最新化が完了したことを確認し、ステップ７４０を経て終了する。レプリカサーバでは最新バージョンの転送７３１により入手した更新データを用いて、ステップ７７５でレプリカの最新化処理を行い、ステップ７６１へ進む。ステップ７６２にて最新化処理終了を応答７７１して終了する。

【００３５】次に、本発明とスヌープキャッシュにおけるキャッシュコヒーレンシーの保証方式との違いについて説明する。スヌープキャッシュでは、データ更新元は

プロセッサであり、バスを監視するのは更新元と同様の機能を持つ他のプロセッサであり、更新対象はメモリであって、前記プロセッサとは異なるレベルの実体である。一方、本発明ではファイルの更新元はクライアントであり、通信路を監視するのはレプリカサーバであり、更新されるのはレプリカサーバと同列の機能を持つファイルサーバにある書き込み可能レプリカである。即ち、監視を行うのが更新要求元と同等か、更新対象と同等かという構成上の根本的な違いがある。また、スヌープキャッシュでは、キャッシュ更新は監視の結果取り込まれた更新データを直ちにキャッシュに反映する。一方、本発明では監視の結果取り込んだ更新データは適当な記憶手段に蓄えられ、データ更新が複数回行われた後でも読み出し専用レプリカの最新化を行うことが可能である。即ち、取り込んだデータの反映の方法に大な違いがある。更に、本発明では、データの最新化にあたり、取り込んだ更新データの妥当性をファイルサーバとの通信を通じて確認することにより、信頼性の高い、レプリケーションを実現している。

【００３６】

【発明の効果】本発明によれば、書き込み可能レプリカの更新と同時にレプリカサーバにもその更新内容が取り込まれるので、読み出し専用レプリカを最新化するにあたり、改めてファイルサーバからレプリカサーバに更新内容を送信する必要がなく、ファイルサーバ、及び通信ネットワークの負荷を増大させることのないレプリケートファイルの最新化が可能となる。更に本発明では、レプリカの更新の際に更新情報の妥当性をファイルサーバとの交信を交えて確認するのでレプリカ間の同一性を高いレベルで保証することが可能であり、信頼性の高いレプリケーションが実現できる。

【図面の簡単な説明】

【図１】本発明の１実施例における適用対象である分散データ処理システムの概要を示す構成図である。

【図２】図１に示されたレプリカサーバの他の実施例の構成を示す構成図である。

【図３】本発明の１実施例におけるレプリカサーバでのデータ取り込み処理を示すフローチャートである。

【図４】本発明の１実施例におけるレプリカ取り込み処理に用いられるデータを示す説明図である。

【図５】本発明の１実施例における妥当性判定用データの構造を示す図である。

【図６】本発明の１実施例におけるファイル、及びレプリカ管理データを示す説明図である。

【図７】本発明の１実施例におけるレプリカ最新化処理を示すフローチャートである。

【符号の説明】

１００　ファイルサーバ（データ処理システム）
１０１　更新データ
１０２　暗合復号キー



１０５、１２５、１４５、１６５　インターフェースボード
１１０　ディスク
１１５　書き込み可能レプリカ
１２０　レプリカサーバ（データ処理システム）
１２１　更新データ
１２２　暗合復号キー
１３０　ディスク
１３５　読み出し専用レプリカ
１４０、１６０　クライアント（データ処理システム）
１８０　通信ネットワーク（ＬＡＮ）
１９０　クライアントとファイルサーバ間の通信
１９１　クライアントとレプリカサーバ間の擬似的な通信
１９２　ファイルサーバとレプリカサーバの間の通信
２００　ファイルサーバと通信ネットワークを接続するインターフェースボード
２１０　レプリカサーバと通信ネットワークを接続する第１のインターフェースボード
２３０　レプリカサーバと通信ネットワークを接続する第２のインターフェースボード
２５０　第２のインターフェースボードからの割込みを処理するプロセッサ
２６０　第１のインターフェースボードからの割込みを処理するプロセッサ
２７０及び２８０　割込み信号
２９０　クライアントとファイルサーバ間の通信
２９１　第２のインターフェースボードクライアントとレプリカサーバ間の擬似的な通信
４００、４２０、４４０　通信情報及びその詳細情報
４０２　宛先アドレス
４２２　宛先識別子
４４１　要求サービス指定子
４４２　ファイル識別子
４４３　更新箇所の相対アドレス
４４４　更新データの長さ
４４５　新たな更新データ
５００　妥当性判定用データ
５１０、５２０、５３０　各回目の更新に伴う冗長データ
６００　ファイル管理エントリ
６０１　ファイル識別子
６２０　レプリカ管理エントリ
６２１　レプリカ識別子
６２３　レプリカの存在アドレス
６２４　レプリカの現バージョン
６２５　レプリカの更新回数
６２６　レプリカ更新データの妥当性判定用データ
７００　ファイルサーバにおけるレプリカ最新化処理
７５０　レプリカサーバにおけるレプリカ最新化処理

15

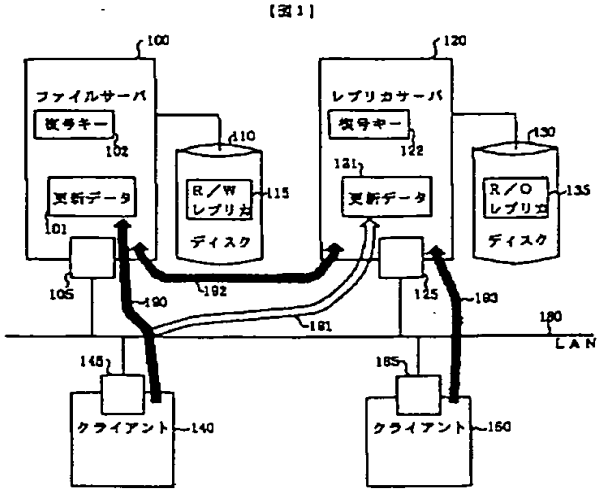
７２０　最新化要求に伴う通信データ

７１１、７３１　ファイルサーバからレプリカサーバへの通信

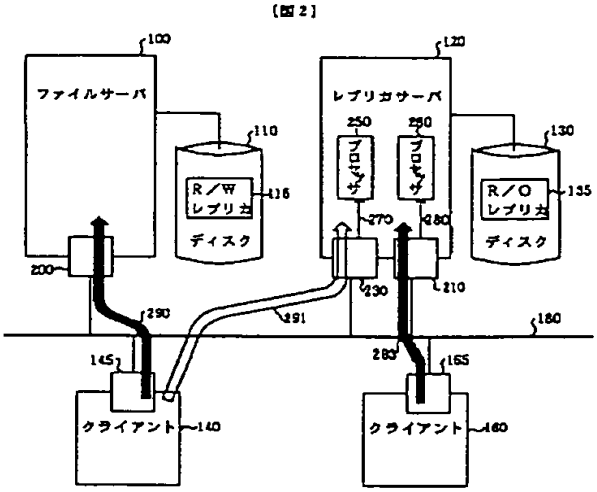
16

７７１、７７２　レプリカサーバからファイルサーバへの通信

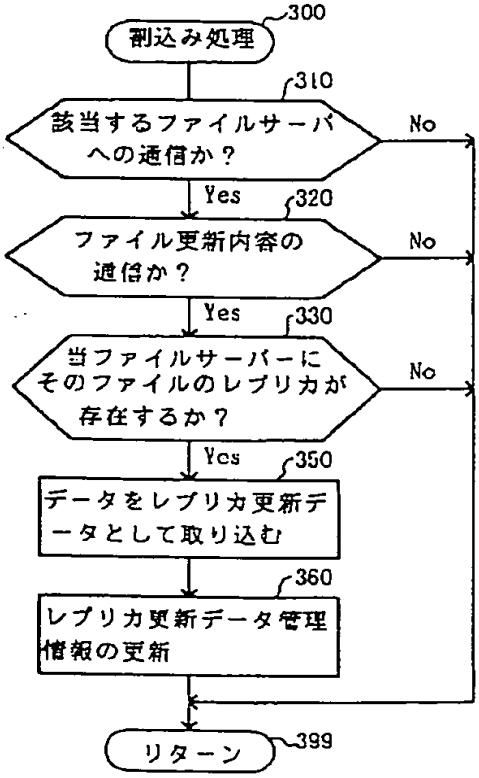
【図１】



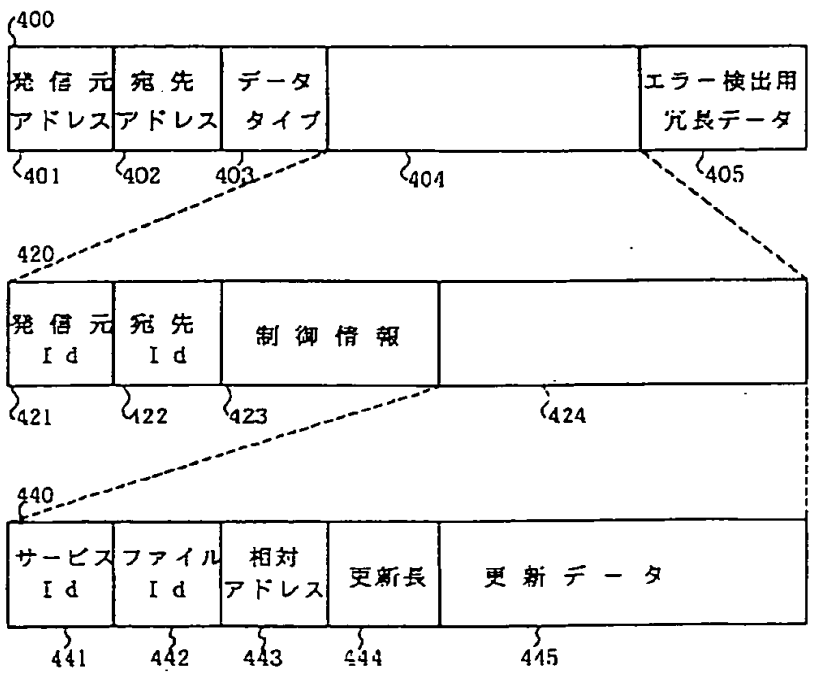

【図２】




【図３】

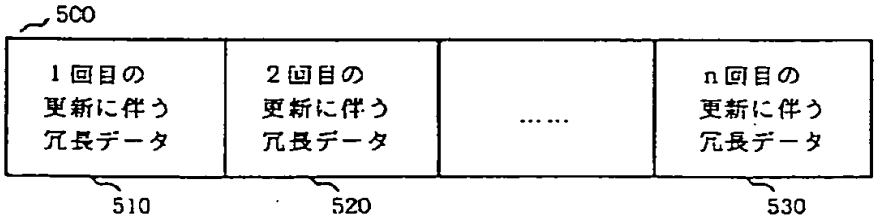



【図４】



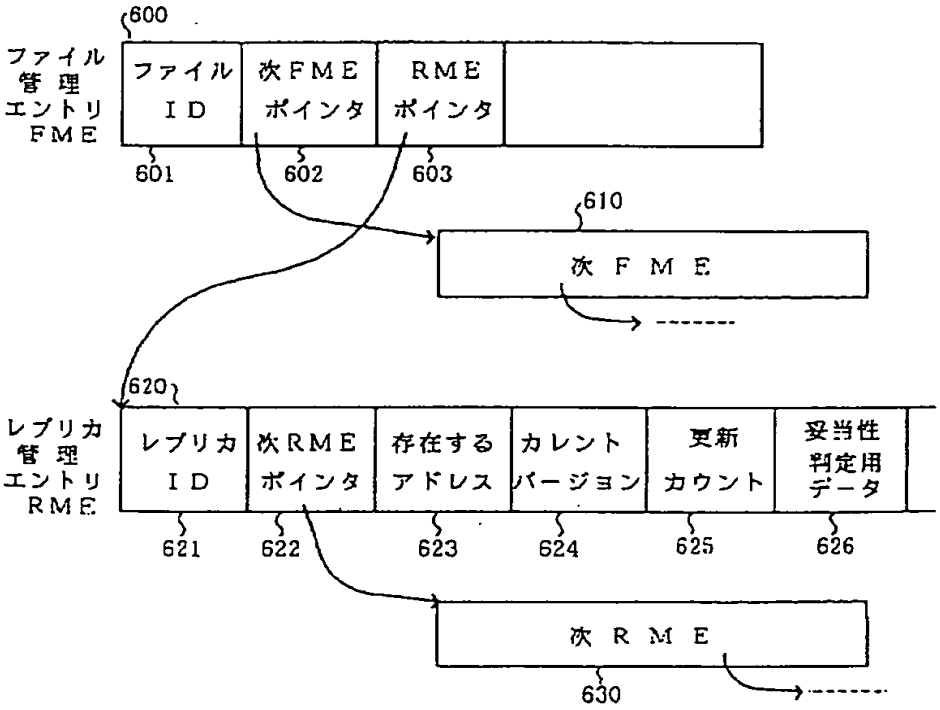
【図5】

【図5】


500
1回目の
更新に伴う
冗長データ
2回目の
更新に伴う
冗長データ
……
n回目の
更新に伴う
冗長データ
510
520
530

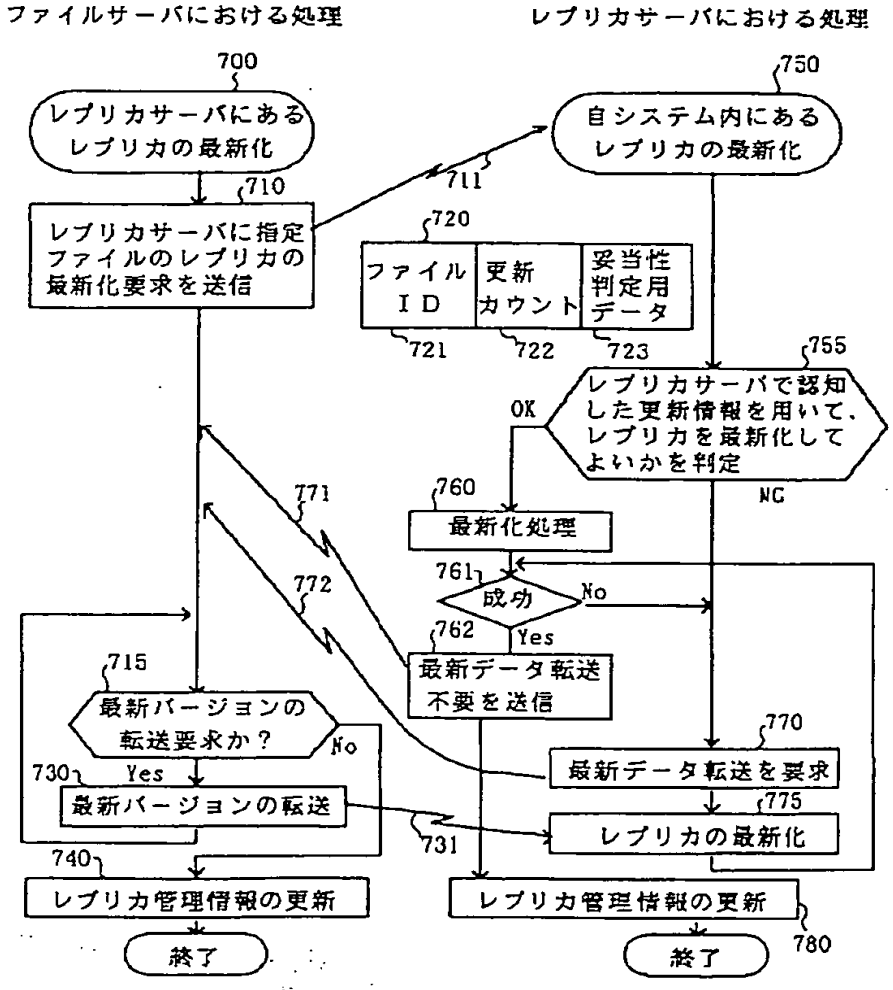
【図6】

【図6】


600
ファイル
管理
エントリ
FME
ファイル
ID
次FME
ポインタ
RME
ポインタ
601
602
603
610
次 F M E
620
レプリカ
管理
エントリ
RME
レプリカ
ID
次RME
ポインタ
存在する
アドレス
カレント
バージョン
更新
カウント
妥当性
判定用
データ
621
622
623
624
625
626
次 R M E
630

【図７】

【図７】


ファイルサーバにおける処理
レプリカサーバにおける処理
700
レプリカサーバにある
レプリカの最新化
710
レプリカサーバに指定
ファイルのレプリカの
最新化要求を送信
711
750
自システム内にある
レプリカの最新化
720
ファイル
ＩＤ
更新
カウント
妥当性
判定用
データ
721
722
723
755
レプリカサーバで認知
した更新情報を用いて、
レプリカを最新化して
よいかを判定
OK
NG
760
最新化処理
761
成功
No
Yes
762
最新データ転送
不要を送信
771
772
715
最新バージョンの
転送要求か？
No
Yes
730
最新バージョンの転送
731
770
最新データ転送を要求
775
レプリカの最新化
740
レプリカ管理情報の更新
終了
レプリカ管理情報の更新
780
終了